A Comparative Analysis of University Entrance Examinations Using the Construct Comparability Approach[1]

Análisis comparativo de las Pruebas de Acceso a la Universidad bajo el enfoque de comparabilidad de constructo

DOI: 10.4438/1988-592X-RE-2020-388-447

Alejandro Veas
Universidad de Alicante
Isabel Benítez
Universidad Loyola Andalucía
Leandro Navas
Raquel Gilar-Corbí
Universidad de Alicante

Abstract

Evaluation processes are a fundamental tool for training and selecting students. However, there are no empirical studies in Spain that analyse the usefulness of assessment tests to measure academic performance. This study, based on research carried out using the construct comparability approach, conducts a comparative analysis of grades achieved by 6,709 students pertaining to 15 academic subject areas of the university entrance examinations (*Pruebas de Acceso a la Universidad - PAU*) administered in the province of Alicante (Spain). The partial-credit Rasch model is used as an estimation method in which each academic subject is regarded as an instrument item related to the measurement of the academic performance construct. The initial results exhibited unidimensionality, and all academic subject areas fit the model, although there was a lack of discrimination between high- and low-performing students, mainly due to the absence of monotonicity in the scoring categories. The difficulty levels

(1) The present work was supported by the Vice President of Research and Knowledge Transfer of the University of Alicante, with grant reference GRE17-16.

Received: 02-05-2019 Accepted: 29-11-2019

of the academic subjects were found to be appropriate for the skill levels of most students. These results demonstrated the ability of the tests analysed to report on the academic performance of the students who took the tests. In addition, important conclusions are presented regarding improvements in the grading processes, and future research studies are proposed.

Keywords: university entrance examination, educational assessment, academic performance, construct comparability approach, partial-credit Rasch model.

Resumen

Los procesos de evaluación constituyen una herramienta fundamental en el marco de la formación y selección de estudiantes. Sin embargo, no existen estudios empíricos en España que analicen la utilidad de las pruebas de evaluación para medir el rendimiento académico. El presente estudio, en base a las investigaciones realizadas sobre el enfoque de comparabilidad de constructo (*construct comparability approach*), realiza un análisis comparativo de las calificaciones obtenidas de 15 asignaturas de las Pruebas de Acceso a la Universidad (PAU) en la provincia de Alicante, con una muestra de 6709 estudiantes. Se emplea el modelo de Rasch de crédito parcial como método de estimación, considerando cada materia como un ítem de un instrumento relacionado con la medición del constructo rendimiento académico. Los resultados iniciales mostraron el cumplimiento de la unidimensionalidad, así como un ajuste de todas materias al modelo, aunque se apreció una falta de discriminación entre sujetos de alto y bajo rendimiento, debido principalmente a la ausencia de monotonocidad de las categorías de puntuación. Se observa que el nivel de dificultad de las materias se adecúa al nivel habilidad de la mayor parte de los sujetos. En base a estos resultados, se destaca la capacidad de las pruebas analizadas para informar sobre el rendimiento académico de los estudiantes. A su vez, se derivan conclusiones relevantes para la mejora de los procesos de calificación, y se proponen investigaciones futuras.

Palabras clave: Pruebas de Acceso a la Universidad, evaluación educativa, rendimiento académico, enfoque de comparabilidad de constructo, modelo de Rasch de crédito parcial.

Introduction

In recent years, there has been an ongoing boom in the study of academic performance at all levels of education. Some studies analyse the cognitive, motivational, and contextual variables involved at the

Received: 02-05-2019 Accepted: 29-11-2019

predictive or causal levels (Dicke et al., 2018; Valle et al., 2008), while others analyse the quality of the measurement of external performance assessment tests and their associated variables (Martí and Puertas, 2018; Sayans-Jiménez, Vázquez-Cano, and Bernal-Bravo, 2018). In the latter case, it is worth highlighting the progress in research on the design and implementation of internationally standardised tests, such as the TIMMS (Trends in International Mathematics and Science Study), the PIRLS (Progress in International Reading Literacy Study), the IALS (International Assessment of Literacy Survey), and, especially, the PISA (Programme for International Student Assessment). However, it should be noted that in Spain, despite in-depth studies of various variables used in the analysis of the abovementioned tests (Elosua, 2013), there has been scant analysis in recent years of the processes for measuring the quality of the PAU (*Pruebas de Acceso a la Universidad*) university entrance examinations beyond some quantitative analyses of the differences between groups of test subjects in specific areas, or of a local nature (Rodríguez-Menéndez, Inda and Peña-Calvo, 2014; Ruiz et al., 2011).

The PAU examination is the current procedure for students who have earned a Spanish baccalaureate to access university studies in Spain and its territories. This examination supports the university admission of students through a series of core and elective academic subject tests. These tests have different formats depending on the academic subject at hand, including text or image commentaries, (short or long) essays on a specific subject, or problem solving, among others. In addition, these test designs are independent of the country's various Autonomous Communities (students from each Community take the same tests), and they produce grades that, when weighted with upper secondary school transcript grades, are used to calculate a total grade used in the students' applications for admission to the various university degree programmes. This process is based on the regulation established in Organic Law 8/2013, of 9 December, for the improvement of educational quality (LOMCE, 2013), and Royal Decree 412/2014, of 8 June, which establishes the basic regulations for admission to official university degree programmes.

Since the PAU is a key assessment process for the futures of thousands of students, it is imperative to consider the role of evaluative research in the field of education. In this sense, it is necessary to begin with a pragmatic and contextual approach for examining the processes and the results obtained, as well as their use in the methods used by various

Received: 02-05-2019 Accepted: 29-11-2019

organisations (Sondergeld and Koskey, 2011) to ensure the principles of equity and equal opportunity for university admission.

In the quantitative research field, various statistical methods have been applied to investigate the fulfilment of the conditions necessary to ensure an objective measurement of academic performance, as well as the correct design of measurement instruments based on the analysis of the most specific conditions. Noteworthy are the use of added-value models and multilevel models for the analysis of longitudinally measured academic performance (Blanco, González, and Ordóñez, 2009; López-Martín, Kouosmanen, and Gaviria, 2014). More specific to the PAU examination is the notable research by Gaviria (2005), in which he uses various statistical techniques (classical method, ordinary least squares method, multilevel method, average equalisation method, and standard deviation) to analyse the equivalence of baccalaureate grades with PAU exam grades, the latter serving as an anchoring point since it is a standardised test for all students. The results show that non-classical methods produce better results than the classical weighting method and make for a fairer student selection process.

However, while the PAU is useful as a standard set of tests for all students (taking into account the necessary knowledge domain-related differences), it is important to ensure that the tests adequately measure students' various ability levels and that they demonstrate an adequate distribution of difficulty levels.

The scientific literature in the academic certification test field includes relevant research conducted in other countries that attempts to analyse the previously described psychometric properties based on different theoretical models for analysing the comparability of academic results. Particularly noteworthy are the models developed in the United Kingdom; more specifically, the performance comparability approach (Baird, Cresswell, and Newton, 2000), the conventional or sociological comparability approach (William, 1996b), the statistical comparability approach (William, 1996a), and a more recently developed model that improves on the previous ones – the construct comparability approach (Newton, 2005). This last model requires that comparisons of two elements have something in common that serves as the basis for this comparison. Just as two tests can be compared by measuring them on the same scale, in the context of academic score comparisons, we can only compare those that measure a common construct – academic performance, in our case.

Received: 02-05-2019 Accepted: 29-11-2019

Thus, the premise of this approach is as follows (Coe, 2008): two scores pertaining to two students are comparable if the academic performance of both (which corresponds to the same level of the latent construct they share) is measured using the same scoring method. According to this postulation, subject matter difficulty will correspond to a specific level established in the latent variable. That is, one academic subject is considered more difficult than another if, to achieve a specific score, a higher level of performance or ability is necessary (Coe, 2010).

Other studies have also demonstrated the need to analyse the usefulness of academic certification tests designed for student admission and to ensure comparability of results. For example, Hübner, Wagner, Hochweber, Neumann, and Nagengast (2019) showed that the results from two tests taken by students in Germany could lead to a deficient admission process because these tests had not been revised to reflect the educational reforms implemented. Furthermore, Korobko, Glas, Bosker, and Luyten (2008) found that the results achieved by students from the Netherlands were influenced by the academic subjects chosen for assessment, revealing the need to adjust the evaluation procedure to avoid an unfair assessment of higher-performing students.

Thus, considering the relevant research on this topic, a comparability measure must use the scores from academic subject tests as an instrument to measure the validity of the construct. This necessity implies that they must adequately represent the content, have good internal consistency, and demonstrate an appropriate degree of correlation between the latent construct and the variables (the different academic subjects).

This measurement model would be impossible without a clear conceptualisation of the construct – academic performance, in our case. It is important to note that, despite being a widely studied concept, there is no single definition of academic performance in the scientific literature. Given the complexity and multidisciplinary approach of academic performance, most of the working definitions refer to the assessment or evaluation of overall achievement at the primary-school level (Jiménez, 2000). However, our construct refers to the achievement levels attained as measured by the respective evaluation standards for the various academic subjects that comprise the PAU tests. This achievement level is translated into specific scores such that a comparison of the construct can be made if increasing or decreasing a score also means progressing or regressing in the construct that is measured.

The theoretical approach presented in this study adheres to the measurement approach espoused by the Rasch model (Rasch, 1980; Wright and Stone, 1979), which is the best-known example of item response theories (IRT). The Rasch model provides a mathematical model based on the calibration of ordinal data from a common measurement scale and checks for conditions such as unidimensionality, linearity, and monotonicity. In its most basic form, this model establishes that the difficulty of an item and the ability of a human subject can be measured on a common scale and that the probability of a person answering correctly will be conditioned by the difference between his or her ability and the item difficulty. Both measures (ability and difficulty) are examined in logit units, since the model uses a logarithmic scale. The use of a common measurement scale allows the setting of homogeneous intervals, such that the difference between the parameters of item difficulty and student ability indicates the same probability of success throughout the whole scale.

At this level of analysis, the starting point is to treat each academic subject as a specific item with 0-10 scoring intervals that indicate different degrees or categories of success. The partial-credit model (Wright and Masters, 1982) enables us to individually analyse the difficulty of attaining a specific score for each academic subject, per the Rasch methodology. This methodology has been used in the UK to analyse the comparability of the General Certificate of Secondary Education certification tests for 16-year-olds and the General Certificate of Education-Advanced tests for 18-year-olds (Coe, 2008; He, Stockford, and Meadows, 2018). The model formula is as follows:

$$Ln\left(\frac{P_{nij}}{P_{ni(j-1)}}\right) = B_n - D_i F_{ij} = B_n - D_{ij}$$

where:

P_{nij} is the probability that subject n will correctly answer item i in category j;

B_n is the measured ability of subject n;

D_i is the measured difficulty of item i; and

F_{ij} is the adjustment measured for item i in category j relative to category *j-1*, which is the point at which categories *j-1* and j are equally likely in relation to the item measurement (Bond and Fox, 2007).

Received: 02-05-2019 Accepted: 29-11-2019

Thus, our main task in this study is the application of the construct comparability approach, developed over the last few decades in the United Kingdom, to the PAU tests administered in one of Spain's Autonomous Communities. Specifically, the objectives are as follows: 1) to compare the adjustment levels and difficulty parameters of the various subjects and 2) to compare the distribution of difficulty levels to the scores achieved in the different academic subjects across the latent attribute.

Methodology

Sample

The sample consists of most of the students from Alicante Province who took the PAU tests in June 2018. Specifically, test scores were collected from 6,709 students who were tested in the province's two public universities: The University of Alicante and Miguel Hernandez University of Elche. The percentage of women in both universities is approximately 60%. The test scores were obtained from the university regulatory agency of the Valencia regional government.

Instruments

The study used the PAU tests administered in Alicante Province in June 2018. These are the same tests administered in the Autonomous Community's other provinces (Valencia and Castellón). Fifteen academic subjects were selected from the set of core and elective subject tests administered. The selection criterion was a minimum of 600 students tested per academic subject. This criterion is used to ensure greater precision of the estimated parameters (He, Stockford, and Meadows, 2018). Thus, the academic subjects selected were Biology, Castilian Language and Literature, Audio-Visual Culture II, Technical Drawing, Business Economics, Physics, Geography, Art History, History of Spain, History of Philosophy, English, Latin II, Mathematics II, Applied Mathematics for the Social Sciences II, Chemistry, and Valencian Language and Literature.

The tests are scored based on standards previously established by each grading commission. The scoring criteria stem from the maximum possible score for each test question, together with qualitative instructions to reinforce the objectivity of the examiners. These scoring criteria are public and available from the *Generalitat Valenciana* website (http://www.ceice.gva.es/va/web/universidad/examenes-y-criterios-de-correccion-de-convocatoria-ordinaria)

Procedure

The construct comparability approach was applied in this study, which assumes that the grades achieved by students in different academic subjects as part of the overall testing process can be compared with one another.

The partial-credit model was applied using the Winsteps (version 4.4.0) statistical software package (Linacre, 2019) for the joint maximum likelihood estimation (Bond, 2004). In this model, each of the included academic subjects is regarded as an item of a single instrument that is capable of measuring the academic performance construct.

First, per the Rasch model, the model's unidimensionality is measured by analysing the main components of the residual scores. According to Linacre (1998), the eigenvalue obtained by comparing the residuals should not be higher than 2.

The process of estimating the item difficulty (including their respective categories) and student ability levels is iterative and examines the relationship between the probabilities of obtaining a certain score as a function of the student's ability. The maximum likelihood procedure enables us to determine the difficulty value of a certain score that best explains the recorded pattern of performance. Similarly, the ability value can be determined for each individual based on the pattern of difficulty indices. This process is repeated using the ability and difficulty estimates until they converge.

While many statistical models try to fit the model to the data, the opposite occurs in this model. In other words, the data must fit the model to be accepted. This fit can be determined from the residual measurements; that is, from the difference between a student's answer for a given item and the expected answer calculated by the model. There

Received: 02-05-2019 Accepted: 29-11-2019

are two ways to standardise the fit measurements for a particular item or subject (Bond and Fox, 2007):

– *Outfit* is the quadratic average of the residuals, divided by the degrees of freedom. This statistic can be interpreted as an overall measure of whether the answers given to a particular item fit the model.
– *Infit* eliminates the extreme scores that influence the outfit, thereby utilising the residuals of the individuals whose ability levels are in the range closest to the specific item.

The infit and outfit statistics are calculated using quadratic averages as a function of the Pearson's Chi-squared statistical value divided by its degrees of freedom, thus producing a scale with values ranging from 0 to infinity. Values below 1 indicate a better-than-expected fit to the model, while values above 1 indicate a poor fit to the model. Thus, if we have an infit value of 1.40, we can assert that there is 40% more data variability compared to the model prediction, while an outfit of 0.80 indicates that there is 20% less data variability with respect to the model prediction.

Different fit values have been established depending on the purpose of the analysis (Coe et al., 2008; Tan and Yates, 2007). Linacre (2002) suggested that values greater than 2 necessarily indicate a poor model fit, and therefore the analysis does not provide reliable conclusions. For this reason, we used this parametric adjustment value for the test items and students, which aligns with previous studies that have also applied the construct comparability approach (He, Stockford, and Meadows, 2018). In addition, the students' average ability level for the various academic subjects was set to zero so that the parameter estimates could be compared with one another.

Results

First, the overall statistics for the model showed a subject (student) reliability index of 0.74 and a subject (student) separation index of 1.69. These values are considered low and indicate that the set of academic subjects is not sensitive enough to effectively distinguish between high- and low- performing students (Bond and Fox, 2007).

Received: 02-05-2019 Accepted: 29-11-2019

Regarding the model's unidimensionality, the results of the analysis of the main components of the residual scores (Bond and Fox, 2007) reveal a main factor that can explain 51.3% of the variance in the latent attribute. The value of the hypothetical second factor is lower than 2 (Eigenvalue V2 = 1.4), which confirms the model's unidimensionality.

Table I presents the academic subjects in order of difficulty (from highest to lowest), as well as their respective goodness-of-fit indices. An optimal fit of all academic subjects to the model is observed in accordance with the established criteria. The academic subjects with the highest difficulty indices are (in descending order) Chemistry, Geography, and Physics. The academic subjects with the lowest difficulty indices are (in ascending order) History of Spain, Mathematics II, and Economics.

TABLE I. Difficulty indices and goodness-of-fit statistics for the analysed PAU academic subjects

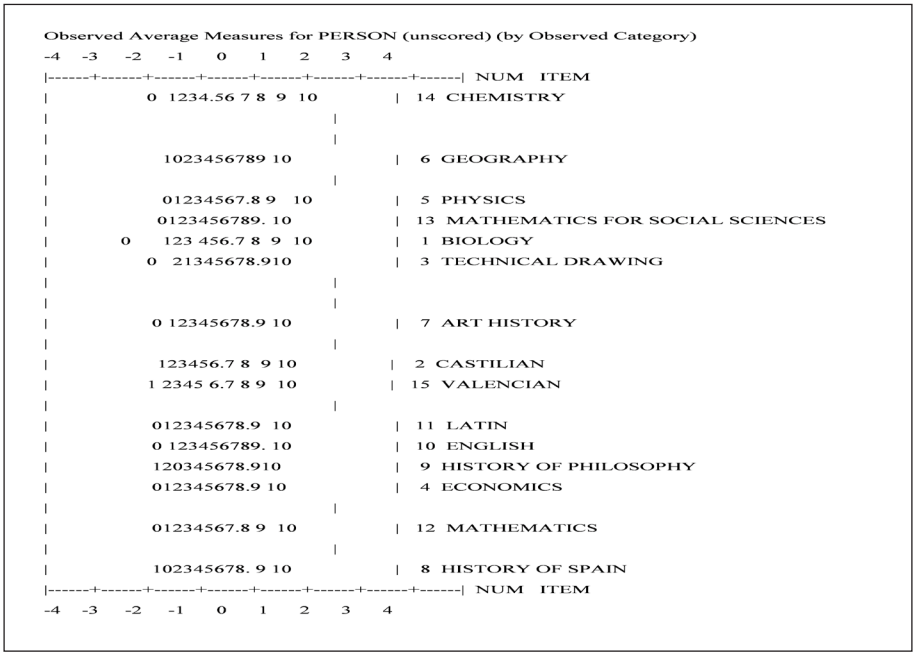
Academic subjects	No. of students	Difficulty	Infit	Outfit	Item-scale correlation
Chemistry	2,179	-0.03	0.92	0.91	0.76
Geography	1,619	-0.16	1.40	1.40	0.62
Physics	1,504	-0.24	1.12	1.12	0.75
Applied Mathematics for the Social Sciences II	2,236	-0.26	1.25	1.25	0.63
Biology	1,738	-0.29	0.90	0.89	0.74
Technical Drawing	697	-0.31	1.51	1.50	0.62
Art History	714	-0.41	1.30	1.29	0.68
Castilian Language and Literature II	6,123	-0.52	0.71	0.74	0.67
Valencian Language and Literature II	4,747	-0.55	0.69	0.71	0.64
Latin	864	-0.63	1.17	1.16	0.69
English	5,960	-0.65	1.17	1.16	0.61
History of Philosophy	778	-0.66	1.37	1.34	0.64
Economics	1,698	-0.67	0.99	0.98	0.70
Mathematics II	3,177	-0.75	1.24	1.19	0.68
History of Spain	6,124	-0.81	0.90	0.92	0.62

Source: Created by the authors based on results obtained using the Winsteps software package

Received: 02-05-2019 Accepted: 29-11-2019

Since the partial-credit model was used, all of the academic subject scores have their own goodness-of-fit indices. In this sense, all of the academic subject scores have an optimal fit, with values ranging from 0.7 to 1.9. However, the score difficulty distribution presented in Figure I shows that the monotonicity criterion was not met for several academic subjects. Thus, no increase in the associated difficulty was detected when moving from one score to the score immediately above it. This situation occurred for the academic subjects of the History of Spain, History of Philosophy, Technical Drawing, and Geography. It was also observed that the scores at the top and bottom of the scale exhibit greater dispersion in the scalar distribution. For example, achieving a score of 10 in Biology is more difficult than achieving a score of 10 in Applied Mathematics for the Social Sciences. Similarly, achieving a score of one in Latin is more difficult than achieving the same score in Valencian Language and Literature II.

FIGURE I. Test score distribution for latent attributes by academic subject

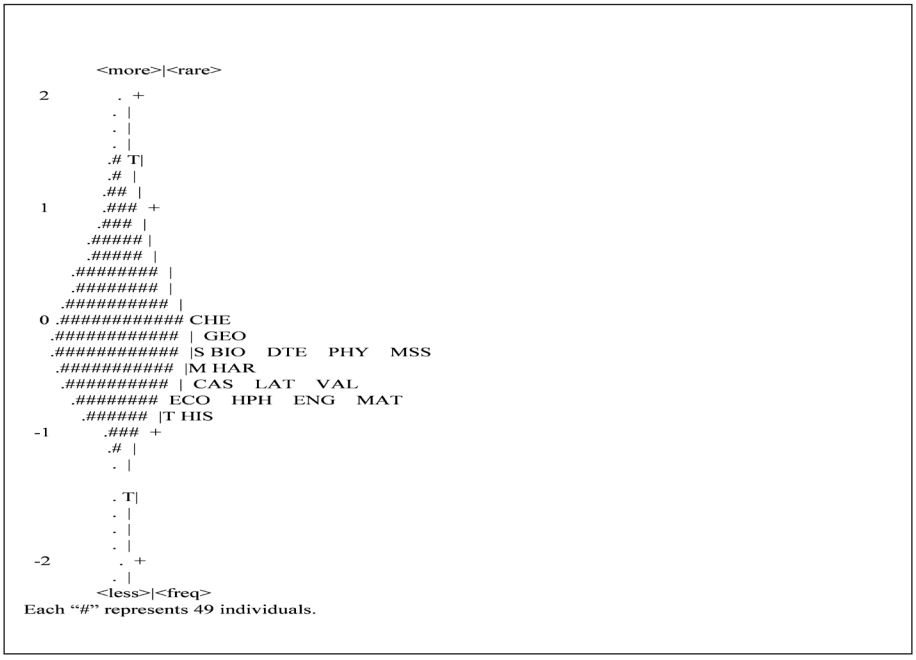


Source: Created by the authors based on results obtained using the Winsteps software package

Received: 02-05-2019 Accepted: 29-11-2019

The "Wright map" presented in Figure II shows the distribution of persons and items along the ranges of ability and difficulty, respectively. Individuals are distributed on the left side of the graph, while items are on the right side. The figure shows that the difficulty of the different academic subjects corresponds to an ability range between logits 0 and -1. This result is positive, since it means that most students are able to take all the exams. There is also a percentage of students between logits 0 and 1, which indicates a higher level of ability and therefore a higher probability of achieving good test scores.

FIGURE II. Person-Item map



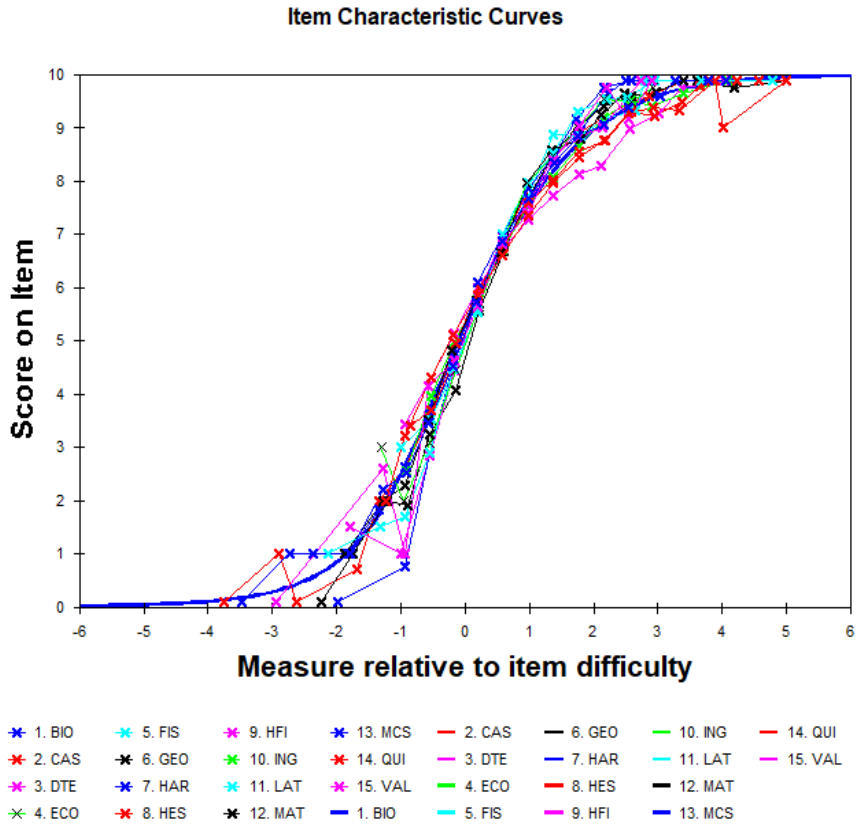
Source: Created by the authors based on results obtained using the Winsteps software package

Lastly, Figure III shows the item characteristic curves (ICC) for all of the academic subjects analysed. These curves reveal the relationships between the expected and observed scores for each item, depending

Received: 02-05-2019 Accepted: 29-11-2019

on the *q* student ability level in the latent attribute. There is good fit when the observed and expected scores overlap between the points and the line, respectively. Specifically, a similar academic subject difficulty pattern is identified in the average scores (between -2 and +2 logits) and in the differences in the maximum and minimum scores (between -4 and -2 logits and between +2 and +2 logits). The easiest academic subjects (based on required ability level) are displayed on the left side of the graph, while the scores that require a higher ability level are on the right.

FIGURE III. Item Characteristic Curves



Received: 02-05-2019 Accepted: 29-11-2019

Discussion and conclusions

The initial analyses demonstrate compliance with the criterion of unidimensionality, which is essential for the application of the model and for establishing a latent construct defined as academic performance within the scope of the PAUs. However, despite the creation of this operational construct, it should not be assumed that a single, overall process exists. The scientific literature indicates that the interpretation of this construct is not clear, since it is not a basis for establishing the specific purposes of each test developed (He, Stockford and Meadows, 2018). Therefore, it is clear that all tests require specific skills, but they all also require global cognitive processes related to the measurement of the construct.

Regarding the first objective, we observed an optimal fit of all the academic subjects analysed, which enables an examination of the invariance properties assumed by the Rasch model (Bond and Fox, 2007). Therefore, the value of this type of estimation lies in the potential for making inferences beyond the sample of students used. At the same time, the academic subject fit allows comparisons with the difficulty parameters obtained from the ability level required to achieve a certain test score. These results lead to a key conclusion in the field of PAU assessment regarding the choice of academic subjects by students, a point that has been widely discussed in the international body of literature (Lamprianou, 2009). Bell et al. (2007) indicate that a student's perceived difficulty of one or more academic subjects may pose a barrier to university admission. Consequently, other academic subjects have higher enrolment rates. Considering the results of this study, such a situation could be occurring with the History of Spain, to the detriment of the History of Philosophy, since students must choose one of these academic subjects, and the former has more than three times as many candidates as the latter.

The analysis of the second objective highlights the need to consider the traditional grading scale used in the PAUs, as the typical monotonicity criterion was not met in the performance tests. The current 10 categories do not adequately discriminate at specific points of the latent attributes. It should be noted that in most countries where comparative analyses of admission test results have been conducted, fewer grade categories are used. Adjusting the scoring system would allow other possible

Received: 02-05-2019 Accepted: 29-11-2019

calculations in the same comparative framework, such as evaluating the relative difficulty of each score by comparing it with the average difficulty of all academic subjects; these differences could be expressed in the logit unit of measurement or in terms of direct scores. This measure would enable observations of how the difficulty of the academic subject scores evolve over successive testing sessions in different academic years (He, Stockford and Meadows, 2018).

Regarding the previous paragraph, the separation rate obtained is low, which affects the fact that the evidence does not discriminate well between students who demonstrate high and low levels of the latent attribute. However, the "Wright map" reveals that all the test difficulty levels are within the students' ability range, so there are adequate degrees of probability of obtaining positive results. The positions of the academic subjects on the scale correspond to a similar distribution of the categories in the latent construct (as seen in the ICC), although some differences can be noted in the distribution of the extremity categories. Once again, these results highlight the need to recodify the categories and improve discrimination by including more students in each high- and low-performance category.

In conclusion, this study aims to initiate an effective analysis in Spain that compares test scores using the construct comparability approach that has been applied in other countries. However, it is important to bear in mind certain limitations that may guide future research on this topic. First, it should be noted that the national data samples used in other countries are much larger. Larger samples enable better estimates because of the higher number of test scores and academic subjects. Conducting this study at a provincial level allowed the use of a construct comparability approach and confirmed the possibility of conducting future studies analogous to those conducted in other countries. In our specific context, this approach enables comparisons between Autonomous Communities to determine appropriate measures for equity. In this sense, the influences of several differential factors must be analysed, such as the individual selection of academic subjects or the effects of educational reforms on testing (Hübner et al., 2019; Korobko, Glas, and Bosker, 2008). Nonetheless, the potential influence of the test graders on the model measurement was not examined in this study and has not yet been explored in the field's scientific literature. However, considering that most PAUs consist of written essay questions, the differences between

Received: 02-05-2019 Accepted: 29-11-2019

test graders regarding task interpretation and evaluation categories (as well as other possible effects such as the halo effect, gender and cultural bias, etc.) can contribute to measurement error and to the validity and fairness of test scoring (Prieto, 2011). The Rasch model can incorporate these types of considerations through an extension of the partial-credit model called the Many-Facet Rasch Measurement.

References

Baird, J., Cresswell, M., y Newton, P. (2000). Would the real gold standard please step forward? *Research Papers in Education, 15*(2), 213-229. https://doi.org/10.1080/026715200402506

Blanco, A., González, C., y Ordóñez, X. G. (2009). Patrones de correlación entre medidas de rendimiento escolar en evaluaciones longitudinales: un estudio de simulación desde un enfoque multinivel. *Revista de Educación, 348*, 195-215.

Bond, T. (2004). Validity and assessment: A Rasch measurement perspective. *Metodología de las Ciencias del Comportamiento, 5*, 179-194.

Bond, T. G., y Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences.* New York, USA: Psychology Press.

Coe, R. (2008). Comparability of GCSE examination in different subjects: an application of the Rasch model. *Oxford Review of Education, 24*(55), 609-636. https://doi.org/10.1080/03054980801970312

Coe, R. (2010). Understanding comparability of examination standards. *Research Papers in Education, 25*(3), 271-284. https://doi.org/10.1080/02671522.2010.498143

Dicke, T., Marsh, H. W., Parker, P. D., Pekrun, R., Guo, J., y Televantou, I. (2018). Effects of school-average achievement on individual self-concept and achievement: Unmasking phantom effects masquerading as true compositional effects. *Journal of Educational Psychology, 110*(8), 1112-1126. https://doi.org/10.1037/edu0000259

Gaviria, J. L. (2005). La equiparación del expediente de Bachillerato en el proceso de selección de alumnos para el acceso a la universidad. *Revista de Educación, 337*, 351-387.

Received: 02-05-2019 Accepted: 29-11-2019

He, Q., Stockford, I., y Meadows, M. (2018). Inter-subject comparability of examination standards in GCSE and GCE in England. *Oxford Review of Education, 44*(4), 494-513. https://doi.org/10.1080/03054985.2018.1430562

Hübner, N., Wagner, W., Hochweber, J., Neumann, M., y Nagengast, B. (2019). Comparing apples and oranges: Curricular intensification reforms can change the meaning of students' grades! *Journal of Educational Psychology.* https://doi.org/10.1037/edu0000351

Jiménez, M. (2000). Competencia social: intervención preventiva en la escuela. *Revista Infancia y Sociedad, 24*, 21-48.

Korobko, O. B., Glas, C. A., Bosker, R. J., & Luyten, J. W. (2008). Comparing the difficulty of examination subjects with item response theory. *Journal of Educational Measurement*, *45*(2), 139-157.

Lamprianou, I. (2009). Comparability of examination standards between subjects: an international perspective. *Oxford Review of Education, 35*(2), 20-226. https://doi.org/10.1080/02054980802649360

Ley Orgánica 8/2013, de 9 de diciembre, para la Mejora de la Calidad Educativa. *Boletín Oficial del Estado (España)*, 10 de diciembre de 2013, 295, 97.858-97.921.

Linacre, J. M. (1998). Structure in Rasch residuals: Why principal component analysis? *Rasch Measurement Transactions, 12,* 636.

Linacre, J. M. (2019). *WINSTEPS rasch measurement computer program* [version 4.4.0]. Chicago, USA: Winsteps.

López-Martín, E., Kuosmanen, T., y Gaviria, J. L. (2014). Linear and nonlinear growth models for value-added assessment. An application to Spanish primary and secondary schools' progress in reading comprehension. *Educational Assessment, Evaluation and Accountability, 26*(4), 361-391. https://doi.org/10.1007/s11092-014-9194-1

Martí, M. L., y Puertas, R. (2018). Comparativa de la eficiencia educativa de Europa y Asia: TIMMS 2015. *Revista de Educación, 380*, 45-74. https://doi.org10.4438/1988-592X-RE-2017-380-372

Newton, P. E. (2005). Examination standards and the limits of linking. *Assessment in Education, 12*(2), 105-123. https://doi.org/10.1080/09695940500143795

Rasch, G. (1980). *Probabilistic models for intelligence and attainment tests* (Copenhagen, Danish Institute for Educational Research). Expanded edition (1989) with foreword and afterword by B. D. Wright. Chicago, USA: The University of Chicago Press.

Real Decreto 1105/2014, de 26 de diciembre, por el que se establece el currículo básico de la Educación Secundaria Obligatoria y del Bachillerato. *Boletín Oficial del Estado (España)*, 3 de enero de 2015, 3, 169-546.

Real Decreto 412/2014, de 8 de junio, por el que se establece la normativa básica de los procedimientos de admisión a las enseñanzas universitarias oficiales de grado. *Boletín Oficial del Estado (España)*, 7 de junio de 2014, 138, 43307-43323.

Rodríguez-Menéndez, M. C., Inda, M. M., y Peña-Calvo, J. V. (2014). Rendimiento en la PAU y elección de estudios científico-tecnológicos en razón de género. *Revista Española de Orientación y Psicopedagogía, 25*(1), 111-127.

Ruiz, J., Dávila, P., Etxeberría, J., y Sarasua, J. (2011). Pruebas de selectividad en matemáticas en la UPC-EHU. Resultados y opiniones de los profesores. *Revista de Educación, 362*, 217-246.

Sayans-Jiménez, P., Vázquez-Cano, E., y Bernal-Bravo, C. (2018). Influencia de la riqueza familiar en el rendimiento lector del alumnado en PISA. *Revista de Educación, 380*, 129-155. https://doi.org/10.4438/1988-592X-RE-2017-380-375

Sondergeld, T., y Koskey, K. (2011). Evaluating the impact f an urban comprehensive school reform: An illustration of the need for mixed methods. *Studies in Educational Evaluation, 37*, 91-107. https://doi.org/10.1016/j.stueduc.2011.08.001

Tasmanian Qualifications Authority (2007). *How the scaled awards are calculated and used to determine the tertiary entrance score.* Available online at: www.tqa.tas.gov.au/0477

Tognolini, J., y Andrich, D. (1996). Analysis of profiles of students applying for entrance to universities. *Applied Measurement in Education, 9*(4), 323-353. https://doi.org/10.1207/s1532481ame0904_3

Valle, A., Núñez, J. C., Cabanach, R. G., González-Pienda, J. A., Rodríguez, S., Rosário, P., … Muñoz-Cadavid, M. (2008). Self-regulated profiles and academic achievement. *Psicothema, 20*(4), 724-731.

Wiliam, D. (1996a). Meanings and consequences in standard setting. *Assessment in Education, 3*(3), 287-308. https://doi.org/10.1080(0969594960030303

Wiliam, D. (1996b). Standards in examinations: A matter of trust? *The Curriculum Journal, 7*(3), 293-306.

Wright, B. D., y Stone, M. H. (1979). *Best test design*. Chicago, USA: MESA Press.

Received: 02-05-2019 Accepted: 29-11-2019

Contact address: Alejandro Veas Iniesta, Universidad de Alicante, Facultad de educación, departamento de psicología evolutiva y didáctica, Carretera San Vicente del Raspeig, s/n, CP: 03690 E-mail: alejandro.veas@ua.es

Received: 02-05-2019 Accepted: 29-11-2019